Instructions on MapReduce Assignments

1. Example of 'Word Count' In Python

- **Description:** use MapReduce method in Python to count the number of each word that appears in the text file.
- **Text File: words.json(just an example)**

  ["row 1", "today is a nice day"]

  ["row 2", "how is she today"]

  ["row 3", "she is very nice today"]
- **MapReduce Method**
  - Map function is a process for generating original 'word-count' pairs for each row.
  - Reduce function is a process for computing the sum of counts for each word.
- **Implementation:**

(You can imitate the mapper and reducer below to accomplish your assignments.)

wordcountMapper.py

```
#!/usr/bin/env python
import sys
import json
# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # parse the line with json method
    record = json.loads(line)
    key = record[0];
    value = record[1];

    # split the line into words
```

```
    words = value.split()
    for word in words:
        # write the results to STDOUT (standard output);
        print '%s\t%s' % (word, 1)
```

wordcountReducer.py

```
#!/usr/bin/env python
import sys
# maps words to their counts
word2count = {}
# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # convert count (currently a string) to int
    try:
        count = int(count)
        word2count[word] = word2count.get(word, 0) + count
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        pass

# write the results to STDOUT (standard output)
for word in word2count:
    print '%s\t%s' % (word, word2count[word])
```

output:

a *1*

very *1*

is *3*

how *1*

she *2*

day *1*

today *3*

nice *2*

✧ **Tips**

- It doesn't matter if you have no Python or Java programing experience. Either C or C++ programing experience is fine, considering that we will not use advanced syntax in Python or Java.
- The first thing you need to do is to understand the idea of MapReduce, and to take a deep look at wordcountMapper.py and wordcountReducer.py. After that you can simply imitate the method to accomplish the other assignments.

2. Practical guide step by step(蓝色部分需要修改路径)

1. $ hadoop fs –ls / (check the directories in hdfs)
2. $ hadoop fs –mkdir /input (make a new 'input' directory)
3. $ hadoop fs –put /path/to/words.json /input (put the input data file into hdfs)
4. $ hadoop fs –ls / (check the directories again, you can find something different)
5. $chmod +x /path/to/wordcountMapper.py /path/to/wordcountReducer.py (添加执行权限)
6. $ hadoop jar /path/to/ hadoop-streaming-1.2.1.jar
   -mapper /path/to/wordcountMapper.py
   -reducer /path/to/wordcountReducer.py
   -input /input
   -output /output

(this command is for execution, you should use the absolute file path, and the output file should not be existed! 请使用绝对路径， streaming 包在 contrib/streaming 目录下)

7. $ hadoop fs –ls /output (check the files in output file)
8. $ hadoop fs –cat /output/part-00000 (show the map-reduce result)
9. *$ stop-all.sh (Remember to stop hadoop before you turn off the machine!!!)*

(You can use ‘hadoop fs –rmr /output’ to remove the output directory,
for more commands on hdfs, please use ‘hadoop fs --help’)